Learning Human Dynamics with Big Data from Online Social Networks

Lilian Weng
Data Scientist, Dropbox, Inc.

With the advent of the Internet, mobile platforms, online commerce, and social media services, the footprints of human behavior are easily recorded in the digital world, generating data on an extremely large scale [Watts, 2007, Vespignani, 2009, Lazer et al., 2009]. Big data documents online discussion in which we participate, the people with whom we interact, how we buy and download a virtual product, and many other aspects of daily routines. The era of big data has brought in many opportunities as well as challenges for researchers by providing details of our lives in many perspectives: from characterization of human genome sequences to message exchanges between online users, and from mobilization of large population recovered through traffic data to easy tracking of information broadcasting. Questions about intricate and complex collective behavior of humans and resulting phenomena, which were hardly captured in self-reported surveys and interviews, can be better answered through the exploration of big data. We can now track, observe, analyze, model, and predict complex human dynamics.

Motivate by the richness of the records of online communication, my research focuses on online social networks and aims to provide insights into communication dynamics and information diffusion processes. My past research projects cover three aspects of online information spreading: actors who share information, features of spreading information (memes), and the interplays between network structure and diffusion process which provide us the capability of predicting which information going viral and modeling the network evolution.

The key methods for processing and investigating big data involve (i) graph theory for the representation of complex human relationships, (ii) modeling and simulation that can verify the hypothesis by replaying the process with fewer but more fundamental factors, and (iii) various machine learning models. Both classification and clustering algorithms are valuable for extracting clean and organized knowledge out of the mass of data. Maximum-likelihood estimation (MLE) estimates the parameters of a statistical model given the independent observed data and furthermore evaluates how good a proposed model can interpret the real-world data.

Though we have developed a set of methods and tools to explore and learn from the big data. There are still several challenges brought about by the massiveness of big data on human dynamics and social phenomena [Watts, 2007, Lazer et al., 2009, Manovich, 2011, Labrinidis and Jagadish, 2012].

Data sampling. The data used in most existing research is generated from a sampling

process. The collection of data can hardly exhaust the whole history of the entire system, due to the policy, privacy problems, and operational issues. An inappropriate sampling method can lead to strong biases. We have only seen a scant amount of work on data sampling algorithms [Lakhina et al., 2003, Leskovec and Faloutsos, 2006, González-Bailón et al., 2012, Morstatter et al., 2013] and the understanding is still limited and insufficient.

Universality. Another problem with the datasets is that most of them are about a single system or a snapshot of the system. We are in need of investigation into whether these results can be applied to other stages in time or other systems without a strong effect of "blind men feeling the parts of an elephant" [Lazer et al., 2009], thus inducing broader impact. Therefore, more future work is expected to study the longitudinal patterns on data with long history and to compare multiple platforms.

Privacy. So many aspects of human society and individual everyday lives have been recorded in big data. It can be dangerous when people are able to look across data from multiple sources to decipher the trace of an individual user. Knowledge about detailed personal schedules and private information may easily facilitate crimes; i.e. breaking into one's house knowing the whole family is on vacation, or illegal access to personal bank account knowing one's address, birth date, and social security number. The prevention of data-based crime deserves enough attention from researchers.

Open access. Data is crucial in quantitative research. However, when a small set of researchers work on private data and produce results, it is important for external people to replicate, verify, or question the findings due to the lack of access to the data. Although many datasets or systems cannot be open to everyone with consideration of privacy, an open environment for scientists and researchers in different institutes, countries, and fields can greatly enhance the vitality and health of academia, encouraging more innovative and meticulous research.

Gap between online and offline systems. User behavior on the Internet is not a transparent reflection of who they are in the real world, because online behavior is usually well curated and systematically managed [Ellison et al., 2006]. This rises up the question of the gap between online and offline environments, while applying classical sociological theorems to the study of online systems, or extending the implication of findings derived from online big data to offline social movements and events.

References

N. Ellison, R. Heino, and J. Gibbs. Managing impressions online: Self-presentation processes in the online dating environment. *Journal of Computer-Mediated Communication*, 11(2):415–441, 2006.

S. González-Bailón, N. Wang, A. Rivero, J. Borge-Holthoefer, and Y. Moreno. Assessing the bias in communication networks sampled from twitter. 1212.1684, arXiv, 2012.

A. Labrinidis and H. Jagadish. Challenges and opportunities with big data. *Proc. VLDB Endowment*, 5(12):2032–2033, 2012.

A. Lakhina, J. W. Byers, M. Crovella, and P. Xie. Sampling biases in ip topology measurements. *Annual Joint Conf. IEEE Computer and Communications*, 1:332–341, 2003.

D. Lazer, A. Pentland, L. Adamic, S. Aral, A.-L. Barabási, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, T. Jebara, G. King, M. Macy, D. Roy, and M. Van Alstyne. Computational social science. *Science*, 323(5915):721–723, 2009.

J. Leskovec and C. Faloutsos. Sampling from large graphs. In *Proc. ACM SIGKDD Intl. Conf. on Knowledge discovery and data mining (KDD)*, pages 631–636, 2006.

L. Manovich. Trending: The promises and the challenges of big social data. *Debates in the digital humanities*, pages 460–75, 2011.

F. Morstatter, J. Pfeffer, H. Liu, and K. M. Carley. Is the sample good enough? comparing data from twitter's streaming api with twitter's firehose. In *Proc. AAAI Intl. Conf. on Weblogs and social media (ICWSM)*, 2013.

A. Vespignani. Predicting the behavior of techno-social systems. *Science*, 325(5939):425–428, 2009.

D. J. Watts. A twenty-first century science. *Nature*, 445(7127):489–489, 2007.